Detecting gene-gene interactions in GWAS using a Group Lasso approach

V. Stanislas[1], C. Dalmasso[1] and C. Ambroise[1]
[1]LaMME, Université d'Evry Val d'Essonne

Framework

Global context

- Genome Wide Association Studies (GWAS) aim at finding genetic markers (SNPs) that are associated with a phenotype of interest. Recently, research topics have been broaden to detect complex genetic structure as **multiple interactions** between markers, known as **epistasis**.
- Epistasis can as well be analyzed at the SNP or at the **gene level**. In the second case, **dimension reduction methods** can be used to resume SNP markers information at the gene scale.
- Even at the gene level the analysis remains in a **high-dimensional** context and the traditional GWAS analyzes that consist on univariate tests perform poorly. Better achievement can be expected with the use of **penalized regression** models adapted to this context as **LASSO**.

Here we propose an approach that takes into account the group structure of each gene to detect epistasis.

➔ Data structure:

	$X_{1,1}$	...	X_{1,K_1}	...	$X_{M,1}$	...	X_{M,K_M}	Phenotype
Ind_1	1		0		0		1	1
Ind_2	0		0		2		1	0
.	2		1		1		2	1
.	0		1		0		0	0
Ind_i	0		2		1		0	0
	$gene_1$				$gene_M$			

General model

$$\boldsymbol{y} = \underbrace{\sum_{m}\sum_{k_m}\beta_{m,k_m}\boldsymbol{X}_{m,k_m}}_{\text{Main effects}} + \underbrace{\sum_{m,m'}\gamma_{m,m'}\boldsymbol{R}_{m,m'}}_{\text{Interaction effects}} + \boldsymbol{\epsilon}$$

- $\boldsymbol{X}_{m,k_m}$: genotype for the SNP k_m of the gene m,
- $\boldsymbol{R}_{m,m'}$: the interaction variable for the given couple (m, m').

➔ Interaction effects definition: Maximum Epistasis Component (MEC)
For each couple of genes we create an interaction variable that maximizes the criterion : $cor[\boldsymbol{Su}, \boldsymbol{y}]$ with,

- $\boldsymbol{S}$ the matrix of all pairwise SNPs product of the two genes,
- u the weight vector that maximize $cor[\boldsymbol{Su}, \boldsymbol{y}]$.

We then define: $\boldsymbol{R}_{m,m'} = \boldsymbol{Su}$.

➔ Coefficient estimation:
We use a group LASSO regression model with a penalty by gene and a penalty by couple.

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}) = \arg\min_{\boldsymbol{\beta},\boldsymbol{\gamma}} \left(\sum_i (y_i - \boldsymbol{X}_i\boldsymbol{\beta} - \boldsymbol{R}_i\boldsymbol{\gamma})^2 + \lambda \left[\sum_{g=1}^{M} \sqrt{p_g} ||\boldsymbol{\beta}^g||_2 + \sum_{c=1}^{C} \sqrt{p_c} ||\boldsymbol{\gamma}^c||_2 \right] \right)$$

λ selected by cross-validation. P-values for each selected group obtained with the adaptive ridge cleaning approach proposed by Bécu et al. [1]

Results

Simulations

Comparison with other interaction methods:
The variables that represent interaction effects in the model can be define in various ways. Here we compare our MEC approach to others that are respectively based on:

- Principal Component Analysis - **PCA**
- Canonical Correlation Analysis - **ACC**
- Partial Least Square - **PLS**

Phenotype generated in two different ways:

- With interaction effect defined as the product of the first PCA component of each gene
- From the model proposed by Wang et al. [2]

➔ Power of the four methods depending on the r^2

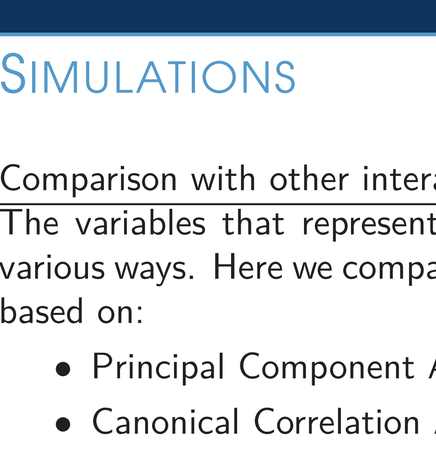

Figure 1: PCA simulation

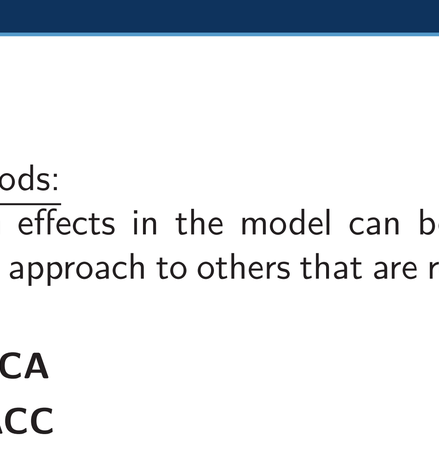

Figure 2: Wang et al. simulation

➔ Effects detected by the different methods

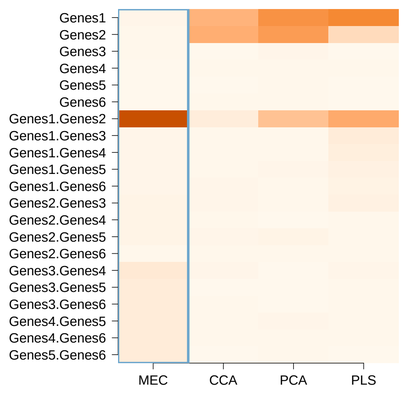

Figure 3: PCA simulation

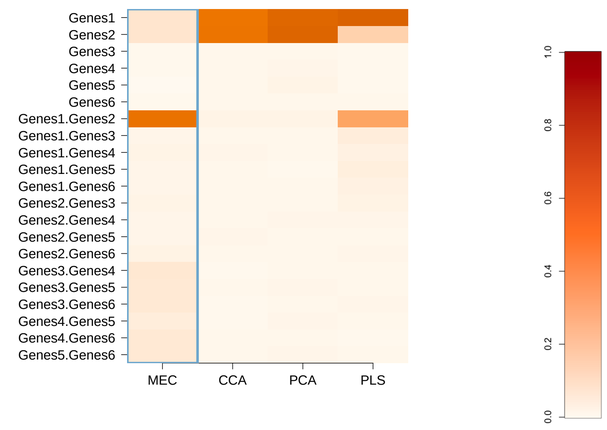

Figure 4: Wang et al. simulation

Theses figures show the ratio of the number of times where each variable was significant on the total number of simulations. Here gene 1 and gene 2 were simulated to have both main and interaction effects with $\beta = \gamma$ and $r^2 = 0.05$.

Real data

We applied our method to a real dataset from a study on ankylosing spondylitis. The data contain 408 cases and 358 controls, and each individual had 116, 513 SNPs genotyped with Immunochip technology. We focused our analysis on a reduced area of 51 genes around the major histocompatibility complex. **One significant interaction** between the genes TNF and LSM2 was identified.

Conclusion

Conclusions

1. By focusing the analysis at the gene level we can facilitate the detection and the interpretation of genetic effects.
2. Our approach gives better performance to detect interactions effects when the r^2 level is low.

References

[1] JM. Bécu, Y. Grandvalet, C. Ambroise, C. Dalmasso, Beyond Support in Two-Stage Variable Selection, arXiv:1505.07281 (2015)

[2] X. Wang, D. Zhang, JY. Tzeng, Pathway-Guided Identification of Gene-Gene Interactions, Annals of Human Genetics (2014)